

2014-11-09 周报

杨哲

本周工作

数据的格式

初步观察了数据，发现了在数据中可能用到的数据字段如下：

Time	Calling_number	Host	Uri	User_agent
20140612_1937	15858185940	m.weibo.cn	m.weibo.cn/pages/100127p213597/weixin?wm=3333_2001&from=timeline&source_type=weixin&uid=2031127437&isappinstalled=0	Mozilla/5.0 (Linux; U; Android 4.3; zh-cn; HTC 8088 Build/JSS15J) AppleWebKit/534.30 (KHTML, like Gecko) Version/4.0 Mobile Safa

本周工作

数据的可利用问题

数据中可利用的用户流量数据非常有限，通过URI仅能探测到用户访问的网站页面，而且通过对几个文件的统计分析发现，用户大部分时间的流量是用于小说，百度搜索，网页新闻，百度知道等内容的查看。统计了几个文件的host访问数如下。

图表标题



本周工作

数据的可利用问题

在分析中发现百度的访问量其实是最高的，但是百度中大部分中新闻及查询难以得到和用户身份相关的信息，我们又对数据进行了部分处理，将数据变为以手机号为单位，在一段时间内对微博的访问情况。示例如下：

```
-----I'm a user , and my phoneNumber is 13486364468*****
-----I'll list only weibo of mine *****
visit host weibo.cn *****
Time: del.20140612_1938_web_0019_00_V0002.dat
URL: weibo.cn/comment/b8yvsa4ua?uid=1648782501&rl=0
UserAgent: Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_1 like Mac OS X)
AppleWebKit/537.51.2 (KHTML, like Gecko) Version/7.0 Mobile/11D201 Safari
Time: del.20140612_1938_web_0019_00_V0002.dat
URL: weibo.cn/u/1648782501
UserAgent: Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_1 like Mac OS X)
AppleWebKit/537.51.2 (KHTML, like Gecko) Version/7.0 Mobile/11D201 Safari
(KHTML, like Gecko) Version/7.0 Mobile/11D201 Safari
```

本周工作

数据的可利用问题

我们分析了上面微博数据的情况，这里只有Get URL的数据，无法查看本账户的登陆情况，只能够查看本账户自己的访问情况，可以看到该账户在何时访问了哪些人，及阅读了的相关内容。

讨论的一种可能的方案

我们在讨论的过程中觉得有种方法可能可以确定微博和手机号码的关联性。

- 1，首先我们生成某用户在使用微博过程中的访问列表
- 2，我们找到在访问过程中一些可能的ID
- 3，我们根据这些ID在微博中爬取相应的微博状态
- 4，将得到的状态和微博列表比对，若符合则为相对应的。

以上方法我们手工找到了一个对应的ID。现在在用算法实现。现在的问题如下：新浪微博的api在今年的时候把获取他人微博的api改掉了，现在只能获取登陆本人的微博列表。所以不能用官方的api，所以只能自己模拟登陆爬数据，但是在大量访问的情况下会被封ip。

下周工作

数据的部分

下周会把本周所剩下的关于新浪微博的算法实现，实际测试一下可用性。

实现部分

下周我们可能会把数据部分暂定，然后实现一些基本的功能。